LLM-Pruner: On the Structural Pruning of Large Language Models

Year: 2023 | Citations: 632 | Authors: Xinyin Ma, Gongfan Fang, Xinchao Wang

---

Abstract

Large language models (LLMs) have shown remarkable capabilities in language understanding and generation. However, such impressive capability typically comes with a substantial model size, which presents significant challenges in both the deployment, inference, and training stages. With LLM being a general-purpose task solver, we explore its compression in a task-agnostic manner, which aims to preserve the multi-task solving and language generation ability of the original LLM. One challenge to achieving this is the enormous size of the training corpus of LLM, which makes both data transfer and model post-training over-burdensome. Thus, we tackle the compression of LLMs within the bound of two constraints: being task-agnostic and minimizing the reliance on the original training dataset. Our method, named LLM-Pruner, adopts structural pruning that selectively removes non-critical coupled structures based on gradient information, maximally preserving the majority of the LLM's functionality. To this end, the performance of pruned models can be efficiently recovered through tuning techniques, LoRA, in merely 3 hours, requiring only 50K data. We validate the LLM-Pruner on three LLMs, including LLaMA, Vicuna, and ChatGLM, and demonstrate that the compressed models still exhibit satisfactory capabilities in zero-shot classification and generation. The code is available at: https://github.com/horseee/LLM-Pruner